SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
August 24, 1997

3. REPORT TYPE AND DATES COVERED
Final; Sept. 25, 1995 - May 24, 1997

4. TITLE AND SUBTITLE
Information Theoretic Investigation of Automatic Object Recognition and Image Fusion

5. FUNDING NUMBERS
DAAH04-95-1-0631

6. AUTHOR(S)
David Cyganski, Richard F. Vaz

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 33785.1-EL

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

19971204 071

13. ABSTRACT *(Maximum 200 words)*

This work involves development of an information theoretic viewpoint for the model-based ATR problem, bringing the problems of ATR and pose estimation (PE) into the framework of signal detection/demodulation theory. Specific goals are to predict maximum achievable ATR/PE performance for a given object/scene scenario, and to quantify the trade-off between performance and system complexity. Progress has been made toward investigating the feasibility of developing useful system performance bounds. Extensions of classical results by Blahut and Fano have shown that tight upper performance bounds can be developed; however, the bounds thus far produced require more information and computation than is practical. Progress has been made toward the long-term goal of designing ATR strategies based directly on information theoretic considerations. A Hilbert Space Decision Tree approach has been developed and tested. This ATR algorithm consists of a series of binary decisions, each of which occurs in the native object image space and maximizes the amount of object discrimination information gained by the decision. In this way, heuristic assumptions regarding the importance of model and image information are avoided. This technique shows great promise for guiding the design of ATR systems.

14. SUBJECT TERMS
ATR, information theory

DTIC QUALITY INSPECTED 4

15. NUMBER IF PAGES
27

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

AN INFORMATION THEORETIC INVESTIGATION OF AUTOMATIC OBJECT RECOGNITION AND IMAGE FUSION

DAVID CYGANSKI AND RICHARD F. VAZ

AUGUST 24, 1997

U.S. ARMY RESEARCH OFFICE
GRANT NO. DAAH 04-95-1-0631

WORCESTER POLYTECHNIC INSTITUTE

Enclosure 2

Contents

List of Figures

List of Tables

1 Problem Statement

This effort involved development of an information theoretic viewpoint for the model-based vision problem. This viewpoint brings the problems of recognition and pose determination into the classical framework of multidimensional signal detection and demodulation theory. It generates a geometric interpretation which allows an intuitive understanding of the behavior of ATR systems, thereby providing insight into the tradeoffs that determine ultimate performance, and suggesting means for optimization. This viewpoint also illustrates why ATR systems that perform well in some cases fail in others, and subsumes the problem sensor data fusion within the same paradigm. Thus, it potentially constitutes a unified framework for the analysis, design, evaluation, and comparison of fusion-based object recognition systems.

The initial goal of this investigation was the development of a unified framework which exploits the parallels between classical communication/information theory and the model-based ATR problem. This initial effort focused on model characterization and ultimate system performance bounds, in addition to development of the analogy between the communication problem and the ATR problem. An additional goal was the characterization of performance tradeoffs between system complexity, system performance, and system failure modes for multi-sensor ATR systems, and development of means for evaluation and comparison of fusion-based ATR systems.

It is anticipated that this will lead to development of guidelines for design and analysis of fusion-based ATR systems. Thus, future work could include investigation into exploiting system performance tradeoffs in order to match ATR techniques to specific sensor modalities and applications, as well as development of model construction and processing techniques for system performance optimization.

2 Summary of Results

In the course of this project the investigators have made significant progress with respect to two directions concomitant to the stated goals of the project:

- investigation of information theoretic performance bounds for ATR performance;
- design of ATR systems based on a direct consideration of information theoretic measures.

The following subsections summarize the more significant results in these two directions. Subsequent sections provide further detail with respect to work conducted since the annual report of January 1997.

2.1 Information Theoretic ATR Performance Bounds

2.1.1 Motivation

The development of Automatic Target Recognition (ATR) systems has had a long history, which has seen successive generations of better performing but increasingly more complex systems. However, little is known about the ultimate performance that yet could be achieved by ATR systems regardless of the type or complexity of the algorithms used.

Information theory, developed in the 1940's by Claude Shannon, has two primary goals, both within the context of communicating a given information source over a channel using coding schemes from a given class. [5]. The first goal is the discovery of the fundamental theoretical limits of achievable performance. The second goal is the development of coding schemes that provide performance that is reasonably close to the optimal performance given by the theory.

The great success of information theory at describing performance bounds in terms of channel characteristics that hold for any communications system together with the fundamental relation between information theory and hypothesis testing (For example, likelihood ratio tests, Stein's lemma, and the Chernoff bound all can be expressed in terms of relative entropy [1]) motivate our application of information theory to the investigation of algorithm independent, theoretical upper bounds on the performance of ATR systems.

In order to investigate the possibility of developing bounds for the absolute performance of ATR systems, we researched existing information theoretic bounds in an attempt to extend their use to the ATR problem. Extensive theoretical and experimental work was done to apply and extend work done by Blahut and Fano to determine not only the usefulness of the bounds but also the computational implications of developing ATR bounds.

We turned to work which had been done by Blahut to determine asymptotic bounds for the performance of binary hypothesis testers, which was based on channel coding theory. Blahut used this analogy between hypothesis testing and channel coding to develop an information theoretic framework within which both upper and lower performance bounds could be developed. We were able to apply this idea to the ATR problem to develop both upper and lower bounds for ATR performance. These bounds were tested numerically, and compared with true ROC operating characteristics developed using a Neyman-Pearson test.

The resulting bounds were not sufficiently tight when compared to an actual system ROC so as to provide useful system characterization.

We then investigated application of Fano's inequality, the principal inequality underlying many of the information theoretic bounds which have been developed for communications applications. The usual application of Fano's inequality results in a bound on total error probability. In the context of ATR however, the Neyman-Pearson viewpoint of detection versus false alarm rate probabilities is of much more usefulness, since it is typically not possible to describe absolute a priori probabilites of the hypotheses or of the costs associated with each type of error. We have produced an extension of the Fano bound which generates a bound on the Neyman-Pearson ROC curve as the envelope of a family of straight-line bounds. This extension produces a tighter upper bound than the Blahut bound on ATR system performance, as assessed by our numerical tests.

It is noteworthy that evaluation of these bounds requires as much information about a system and as much computation as needed to produce the actual ROC curve itself. Thus, application of these bounds is no more computationally practical for application to a complex ATR scenario than direct computation of the ROC.

2.1.2 Information Theoretic ATR System Design

Our other efforts have focused on the design of an ATR system architecture which is directly motivated by an information theoretic posing of the ATR problem. Using the geometric image space model for representation of target instantiations which was presented in the proposal for this work, we have developed an Information Theoretic Decision Tree (ITDT) ATR procedure. This approach frames ATR as a sequence of decisions leading to object identification and pose estimation in which the number of binary decisions is minimized by maximizing the information gained by each decision. These decisions are framed in the native image space defined by the raw object image data, so as to not sacrifice information due to artificially imposed lossy data compression, as is the case with any reduced-data approach to ATR.

Since each decision in the ITDT process seeks to maximize the information gained by that decision in the sense of discrimination between object instantiations, establishment of the decision regions is an NP-complete global optimization problem. However, we have developed a local optimization approach based on gradient descent techniques in which the non-optimality of the decision sacrifices only ultimate speed of processing, and not the ultimate ATR performance.

We have tested this system with respect to recognition of targets with azimuthal freedom, from SAR imagery, embedded in clutter. The results, shown in the sequel, indicate performance levels well above that of other ATR systems for pose-variable targets currently under study at WPI with significantly less processing.

These encouraging preliminary results provide evidence that information theoretic considerations can be used as the basis for ATR system design.

2.1.3 Hierarchical Decision Tree based ATR

Hierarchical search of a database of target exemplars offers the possibility of vast decreases in processing time for a given ATR performance since:

- Hierarchical search for data entities that can be arranged in an ordered binary tree can yield linear processing time growth for exponentially increasing data sets. Hence the much desired goal of linear growth of ATR processing time with numbers of degrees of target pose and environment freedom seems obtainable.

Hierarchical search of a database of target exemplars has often been investigated and rejected by various researchers owing to spurious decision behaviors discovered on extensive testing because typical implementations of hierarchical search:

- pre-suppose that the model imagery forms a totally ordered (sub-linearly searchable), which in general it does not;
- reduce the dimensionality of the imagery by some form of projection to obtain a searchable set with an unavoidable accompanying loss is discrimination power;
- apply hash functions without regard to the unruly effects of image degradation (from sensor noise, environment and target variation) on the hash function output.

2.1.4 Information Theoretic Decision Tree based ATR

During the first portion of this investigation into the implications of information theory with regards to the ultimate performance and bounding of performance for ATR, certain insights arose regarding effectual construction of decision tree based ATR systems:

- Total ordering can be imposed upon the exemplar set by "extension" of the set in which new set members are generated by properly dividing the original exemplar equivalence classes so that a given exemplar may be represented by several equivalence classes;
- the original image space should be divided, and not a projection of the true image space, to retain all relevant information;
- the division of the space should be driven by maximization of an information theoretic measure (to be described below) to minimize the number of decisions and hence the amount of computation;
- the penalty for approximation in the division of the decision space in this case is increased computation but not loss of performance.

2.2 Discussion of the ITDT Development

The following slide depicts the placement of points in a 3-D decision space generated from the intensity of three pixels in a sequence of 14 target exemplar images. Each of the points represents the center of density function which models the stochastic behavior of the target, sensor and environment.

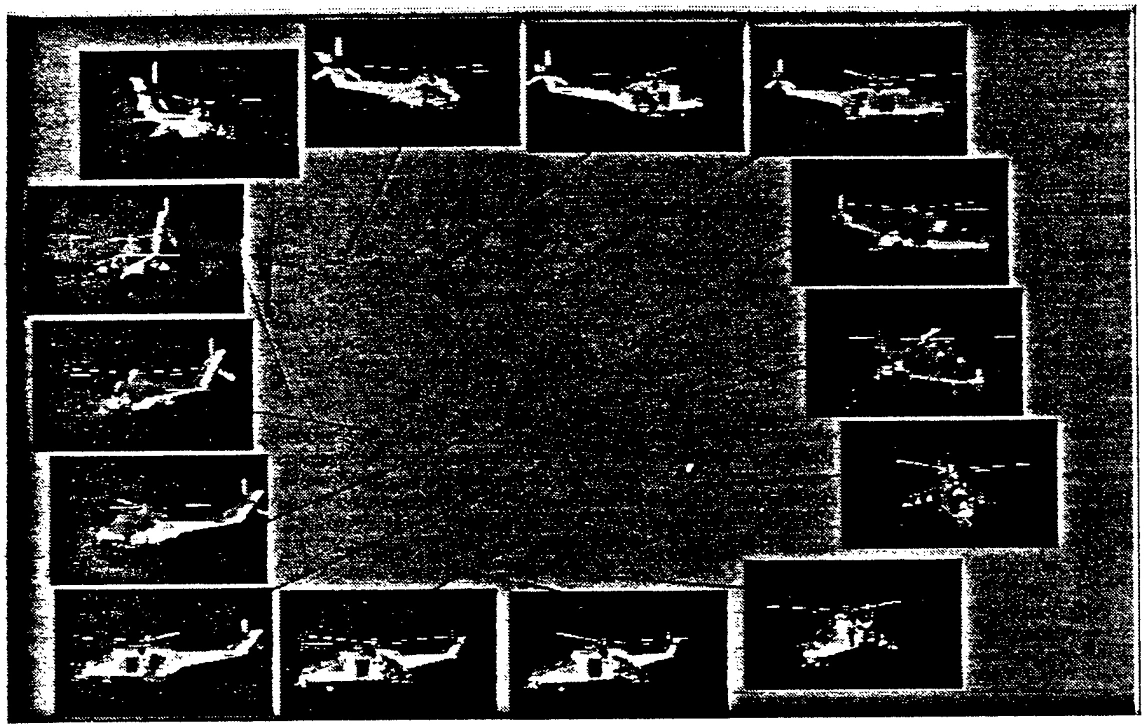

Figure 1: *Depiction of a 3-D decision space occupied by points derived from image data.*

When considering the question of how to divide the problem of image recognition into a set of simple, hierarchical steps, we intuitively are led to the notion of successively splitting the space by a set of planes. Choosing these planes to segregate exemplar points into approximately equal subsets yields a binary search tree that can be navigated in a $log_2(M)$ steps given M exemplars. In later portions of this report we will explore the mathematical basis and optimization of this division operation. At this point we will simply consider the overall notion and its implications.

Now consider the placement of a decision plane that divides this space into regions containing approximately half the number of image points. The process of dividing this space must take into account the probability density function that describes the distribution of points that would be obtained due to the stochastic nature of the target, environment and sensor system. For illustrative purposes in this example, we will show the extent of the 1 sigma radius of the density functions as the surface of a translucent ball in the 3-space as shown in Figure 2. A division of this space is depicted in Figure 3.

It is inevitable that a given decision plane will cut through many (if not all) of the distributions associated with the exemplar points. More importantly, we must consider that the exemplars are segregated by a true equivalence class structure that is generated by evaluation of the classic likelihood function associated with the given density functions. It is the event in which a decision plane cuts through the boundaries of the true equivalence classes that gives rise to an irrecoverable loss of ATR performance. That is, upon segregating exemplars on this basis of such a division, we have placed a region of the decision space associated with a given exemplar on a side of the decision tree that can never yield that exemplar as a best match.

Figure 2: *Depiction of the 3-D decision space in which the stochastic nature of the imagery is represented by spheres of radius equal to one standard deviation of the associated probability density.*

A means to circumvent this problem is to place such exemplars in both subsets generated by the division. As a result, certain exemplars will appear in several exemplar subsets associated with the decision tree branches, implicitly generating the equivalence class extension needed to generate a totally ordered set of classes. It is by applying a simplified scheme derived from this notion that the implementation tested below demonstrated freedom from the poor early decision problem often found with hierarchical systems.

2.2.1 Decision hyperplane placement

In the more detailed treatment to follow, it will be argued that the placement of decision hyperplanes that cut through the image space and divide it into equivalence regions, and the exemplars into subsets, is properly determined by the maximization of the information measure:

$$I(\{p_i\};Y) = \mathcal{H}(\{p_i\}) + \mathcal{H}(Y) - \mathcal{H}(\{p_i\},Y),$$

where $\{p_i\}$ represents the joint density function (determined by the exemplars and overall statistical system behavior) and Y is the binary valued function over the image space determined by the hyperplane position.

This optimization maximizes the information gained about the identity of any point in that space which respect to exemplar equivalence by means of that decision. Obviously we can gain at most 1 bit with each decision. Gaining one bit per decision would allow determination of membership of any image in an exemplar decision set in the minimum, logarithmic, time.

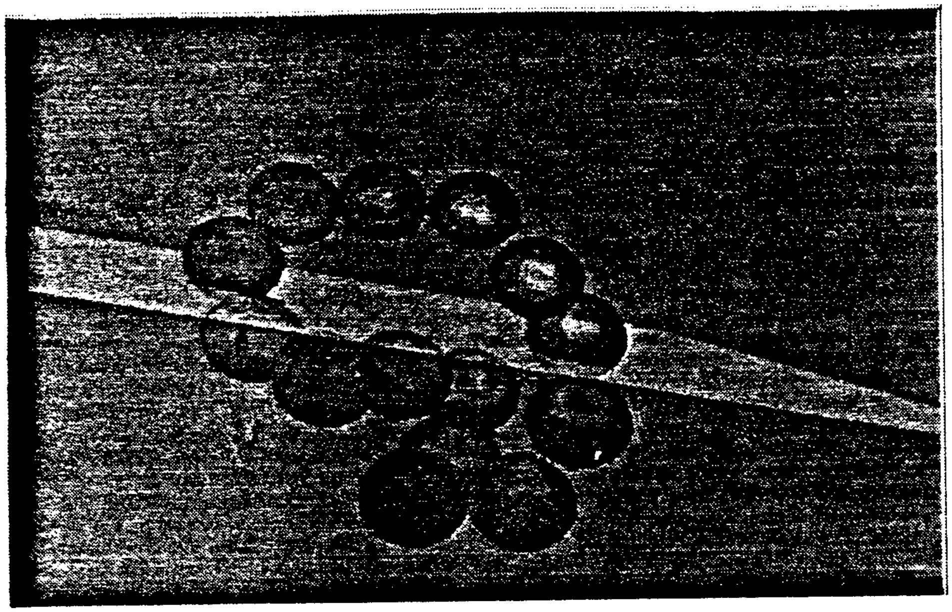

Figure 3: *Depiction of the division of the 3-D decision space such that each half-space is occupied by approximately one-half of the exemplar image points.*

2.2.2 Significant Attributes of the ITDT ATR

The Information Theoretic Decision Tree ATR possesses the following attributes:

- The on-line ITDT decision process grows approximately logarithmically with exemplar size. With the 106 exemplar based system, to be demonstrated, only 7 decisions are needed to establish the nearest exemplar to a given input image.
- The off-line ITDT construction process is very computationally intensive. Computation of a 106 target exemplar based system with the simplest additive i.i.d. statistical description of sensor/target stochastic behavior requires 2 hours on a 200 MHz workstation. This will increase linearly with exemplar size.
- The off-line computation involves the approximate solution of a global optimization problem involving mutual information measures computed from exemplar statistics. As in any attempt at global-optimization with bounded computational resources, the solution is only a local optimum. The beauty of the ITDT approach is that the non-global nature of the solution has no effect on the performance of the ATR, it only potentially increases the number of decisions that must be executed. (In all of our 1-DOF tests, to date, we have not seen more than a single additional decision level as a result of this non-optimality.)

2.3 ITDT Performance Results

We have constructed a prototype that demonstrates the new approach for 1-DOF problems and have tested this system with the same SAR data sets used in the ARO and DARPA

sponsored project that has led to the development of two other ATR systems for variable target pose recognition, LSD/DOA and PERFORM. As we will compare performance with these two systems, a brief description of each follows:

2.3.1 LSD/DOA ATR architecture

Developed by WPI under a grant from the ARO and DARPA the LSD/DOA (Linear Signal Decomposition / Direction of Arrival) ATR is a model-based target identification and pose estimation algorithm which reduces a pose dependent object description into a small, fixed, non-linear filter system for ATR and estimation of pose parameters for targets with unknown pose [7, 8].

The technique factors the ATR problem into two components: an off-line analytic model construction process and an on-line direct determination of object identity and pose. This architecture allows fast, non-iterative object recognition by shifting the computational burden of ATR to the off-line filter construction.

While for classic ATR systems the computational burden lies in the on-line (real-time) processing, involving repeated interrogation of the object model, for LSD/DOA the burden lies in the off-line (pre-mission) reciprocal-basis-set (RBS) construction process. This basis set is constructed in such a way that linear projection of an acquired object image onto the basis images yields a set of harmonically related multidimensional complex samples. From these samples object pose parameters may be estimated using direction-of-arrival (DOA) techniques. Thus, LSD/DOA provides a means to perform ATR with any object type with parameterizable behavior: opaque, refractive, rigid, articulated, etc.

Furthermore, the decomposition of the recognition system into a variable size system of linear filters that generate the data for non-linear estimation, allows configuration that can balance computational complexity and data storage with performance requirements.

2.3.2 PERFORM ATR architecture

The Partial Evidence Reconstruction From Object Restricted Measures (PERFORM) algorithm was derived from LSD/DOA for the purposes of demonstrating recognition of partially obscured targets [9, 10]. Its derivation is based upon the observation that by confining an RBS support region to the interior of the object support intersection over the full pose range, we can obtain information about an object without introduction of clutter induced corruption without an accompanying loss of object information and hence discrimination power.

By using several such target embedded estimators we can develop a set of partial object evidence which can be assembled into a single object pose and identity hypothesis. A direct benefit is greatly increased robustness to object obscuration in so much as certain cover estimators may be altogether unaffected by partial obscuration.

2.3.3 Performance of the ITDT ATR

In the prototype ITDT implementation, a 1369 dimensional space was populated by 106 exemplar points with a distribution model based on the assumption of additive i.i.d. Gaussian noise. A 7 level decision tree is constructed from this data.

The system was tested using SAR imagery. The target data was generated from spotlight SAR phase history files provided by Wright Laboratories, Wright-Patterson AFB. The test results presented below are based on data from a Soviet T72 tank. The target exemplars were generated from L band data, a 10 degree elevation angle, and HH and VV polarization data which were used to form a single, polarimetrically whitened image. The background data was taken from the ADTS data set obtained from Lincoln Laboratories. The images used were polarimetrically whitened SAR images depicting terrain in Stockbridge NY at 1 ft. by 1 ft. resolution. The final test images were obtained by overlaying the targets onto the clutter backgrounds by masking out a region of the clutter corresponding to the convex hull of the brightest target pixels and inserting the target image into the masked area.

The original data set allowed for the construction of 318 poses of the target at equi-spaced angles from 0 to 360 degrees. Throughout the following discussion we used only a subset of this set corresponding to 106 images or 53 (every other image of the 106) representing poses from pose angle 0 to 120 degrees. This restricted range was used to better expose the behavior of the ITDT as a pose estimator in noise without the apparent large errors that are induced by selection of estimates opposed to the correct angle by 180 degrees due to the near perfect symmetry of the target at certain poses. The 53 image set was used to test the effects of less complete training sets.

In Table 1 we show the performance of the ITDT system with respect to error of estimation of azimuthal pose angles of the test targets under various conditions. The first entry describes an evaluation which simply tests the sensitivity of the system to clutter backgrounds. That is, the test images were constructed by placing the 106 training images on a large number of clutter backgrounds. If the algorithm was totally insensitive to the clutter background, the error would be zero in this case as no other perturbations are present. The small 0.67 degree std. dev. reveals a very small perturbation of the estimate.

The second entry in the table describes pose estimation performance in the case of an ITDT constructed from the 53 image subset of target images and tested with the full 106 image set placed on various clutter backgrounds. Furthermore, the test images were corrupted by the addition of simulated SAR speckle noise. Again the performance is quite good with a pose estimate std. dev. of 0.86 degrees.

Finally the third entry tests the improvement that can be obtained by increasing the size of the training set. Here the ITDT is constructed from all 106 target images and tested with the full 106 image set placed on various clutter backgrounds. Again, the test images were corrupted by the addition of simulated SAR speckle noise. Again the performance is quite good with a pose estimate std. dev. of 0.66 degrees.

The prototype successfully demonstrates logarithmic processing time growth with increasing exemplar set size and a performance that exceeds all of our previous ATR systems as shown in the following ROC curve. To compare the performance for a given amount of processing, we need to make the following observations:

- The number of floating point operations associated with the ITDT algorithm for each image tested is approximately equal to $7N$ where N is the number of pixels in the region of interest and the constant, 7, obtains from the number of decisions made.
- The number of floating point operations associated with the LSD/DOA algorithm for each image tested is approximately equal to $15N$ where N is the number of pixels in the

Decision Tree	Test Suite	error mean	std. deviation
t72_L_10_pwf.dB.156.f106.DT	t72_f106.onbg.suite	0.28 degrees	0.67 degrees
106 image (0..120 deg.) DT tested with 106 unspeckled images(0..120 deg.) on clutter			
t72_156.f106e53.DT	t72.157.f106.onbg.suite	0.76 degrees	0.86 degrees
53 image (0..120 deg.) DT (2.3 deg. spacing) tested with 106 speckled images(0..120 deg.) on clutter			
t72_L_10_pwf.dB.156.f106.DT	t72.157.f106.onbg.suite	0.28 degrees	0.66 degrees
106 image (0..120 deg.) DT (1.14 deg. spacing) tested with 106 speckled images(0..120 deg.) on clutter			

Table 1: Performance of ITDT with respect to estimating the azimuthal pose angle of targets under various conditions.

region of interest and the constant, 15, obtains from the number linear filter operations (4) and number of metric evaluations (11).

- The number of floating point operations associated with the ITDT algorithm for each image tested is approximately equal to $15N$ where N is the number of pixels in the region of interest and the constant, 15, is given as for LSD/DOA.

Thus we see that at less than half the processing load, the ITDT achieves performance dramatically in excess of the other methods at the lowest false alarm rates.

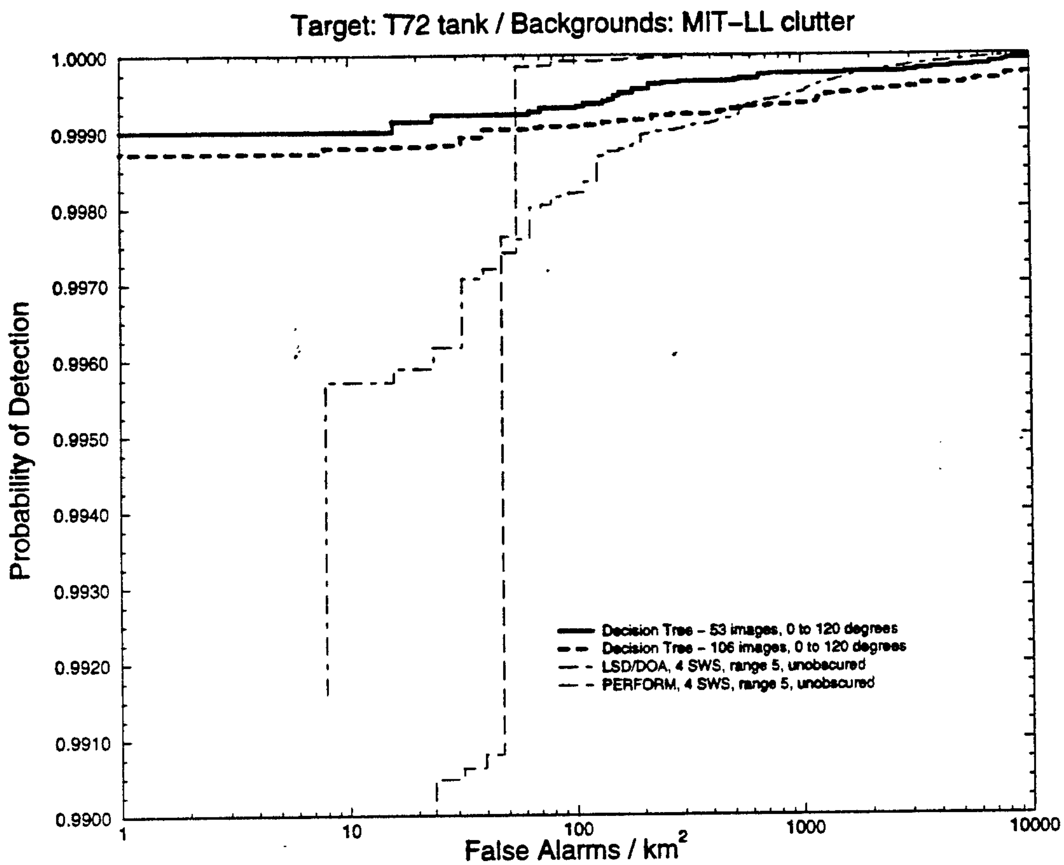

Figure 4: *ROC curves for the ITDT, LSD/DOA and PERFORM ATR systems.*

3 Derivation of Information Theoretic ATR Performance Bounds

We develop in this section a set of ATR Performance bounds based upon the application of information theory to decision problems. After discussing the relevance of information theoretic bounds on the ATR problem, and reviewing certain well known results, an example application is examined. The example demonstrates that the Blahut bound is extremely loose for even a simple problem. This leads to the derivation of a new tighter bound based on Fano's inequality which we call the Fano-Envelope bound.

3.1 Approach

We consider the problem of correctly recognizing a single target from a finite set of possibilities. Define a discrete random variable, X, with alphabet $\mathcal{X} = \{c_1, c_2, c_3, \cdots c_N\}$, corre-

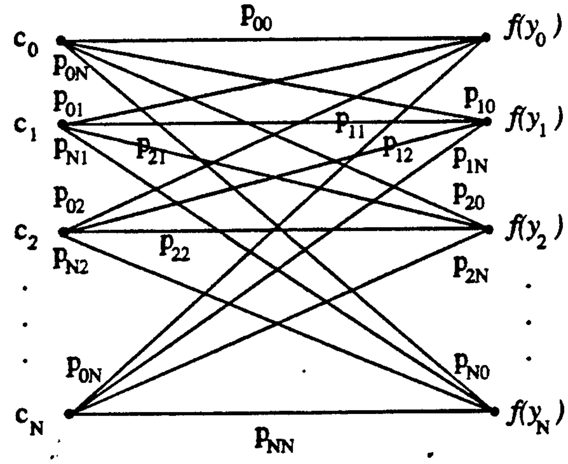

Figure 5: An N-ary target recognition problem with input alphabet equal to the set of possible targets and output alphabet equal to the set of decisions made by the ATR.

sponding to the set of all target classes under consideration, let $p_X(c_i)$ be the probability mass function of X. An ATR makes an observation Y, which is another discrete random variable with alphabet $\mathcal{Y} = \{y_1, y_2, y_3, \cdots y_N\}$, and described by the conditional probabilities $p_Y(y_i \mid c_j)$. Finally, from Y an estimate of X, $\hat{X} = f(Y)$, is made. This simplest form of the target recognition process is a random mapping from the set of possible targets onto itself, and is shown in Figure 5.

The receiver operator characteristic (ROC) diagram plots the probability of detection against the probability of false alarm for binary hypothesis testing problems. The curve is generated by parametrically varying the threshold applied to the decision metric, which is the only statistic available to distinguish between the two hypotheses.

In the binary hypothesis testing problem, $\mathcal{X} = \{c_0, c_1\}$ and we denote the probability of making a Type I error (false alarm, or false positive) as $P_{10} \equiv p(y_1 \mid c_0)$ and the probability of making a Type II error (missed detection, or false negative) as $P_{01} \equiv p(y_0 \mid c_1)$. For notational convenience, we let $\pi_0 = p_X(c_0)$ and $\pi_1 = p_X(c_1)$, and recall that $\pi_1 = 1 - \pi_0$.

We will now address the question of how to combine information theoretic concepts and translate them into a form compatible with the ROC diagram evaluation of ATR performance. In particular, we will examine Fano's inequality from information theory, generate a performance bound on a ROC diagram for the binary hypothesis testing problem, and compare it to another ROC performance bound derived from information theoretic concepts by Blahut [2].

3.2 Fano's Inequality

Suppose we know a random variable Y and wish to estimate the value of a correlated random variable X. Fano's inequality relates the probability of error in estimating the random variable X to its conditional entropy $\mathcal{H}(X \mid Y)$.

Theorem. Fano's Inequality [1]. Let X, Y be random variables with respective alphabets $\mathcal{X}$ and $\mathcal{Y}$. Let Y be related to X through the conditional probabilities $p(y \mid x)$. From Y, an

estimate of X $\hat{X} = f(Y)$ is produced. If we define $E : \mathcal{X} \times \mathcal{X} \longrightarrow \{0,1\}$ to be a binary error random variable such that

$$E(x, f(y) = x) = 0$$
$$E(x, f(y) \neq x) = 1$$

with probability $P_e \equiv Pr(E = 1)$ and entropy

$$\mathcal{H}(E) \equiv \mathcal{H}(P_e) = -P_e \log_2(P_e) - (1 - P_e)\log_2(1 - P_e). \tag{1}$$

It is the case then that

$$P_e \log_2(|\mathcal{X}| - 1) + \mathcal{H}(P_e) \geq \mathcal{H}(X \mid Y), \tag{2}$$

where $|\mathcal{X}|$ is the number of elements in the alphabet of the random variable X, and the conditional entropy is defined to be

$$\mathcal{H}(X \mid Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(xy) \log_2 p(x \mid y). \tag{3}$$

In the binary hypothesis testing case, $|\mathcal{X}| = 2$, so Fano's inequality reduces to

$$\mathcal{H}(P_e) = -P_e \log_2(P_e) - (1 - P_e)\log_2(1 - P_e) \geq \mathcal{H}(X \mid Y). \tag{4}$$

3.3 Blahut's Bound

The Blahut bound on the probabilities of Type I and Type II errors, P_{01}, P_{10}, is built on the relative entropy (discrimination, Kullback-Leibler distance) for a pair of discrete random variables X and Y with respective distributions $p(x_k)$ and $p(y_k)$ [2]. Specifically, if $P_{10} \leq e^{-nr}$, then

$$P_{01} > 1 - \frac{4\sigma^2}{n(r-c)} - e^{-n(r-c)/2} \tag{5}$$

where

$$\sigma^2 = \sum_k p(y_k) \left(\log \frac{p(y_k)}{p(x_k)}\right)^2 - \left(\sum_k p(y_k) \log \frac{p(y_k)}{p(x_k)}\right)^2, \tag{6}$$

$$c = \sum_k p(y_k) \log \frac{p(y_k)}{p(x_k)} \tag{7}$$

is the relative entropy, $r > c$ is the relative entropy between a dummy distribution and $p(x_k)$, which is varied to produce samples for the bound, and n is the number of independent measurements. Blahut's bound has its roots in the block coding converse to the coding theorem for noisy channels in information theory [6]. There r corresponds to the transmission rate of a source, c the capacity of a discrete memoryless channel, and n the block length.

3.4 Performance Bounds on a ROC curve

A ROC curve plots the probability of detecting a target, $1 - P_{01}$, against the probability of false alarm, P_{10}. The total error probability is related to these conditional probabilities,

$$P_e = P_{01}\pi_1 + P_{10}\pi_0. \tag{8}$$

This equation can be rewritten as,

$$P_{01} = -P_{10}\frac{\pi_0}{\pi_1} + \frac{P_e}{\pi_1}, \tag{9}$$

a relation between P_{01} and P_{10} corresponding to a curve on a ROC diagram.

It is apparent from this expression that for given prior probabilities (π_0, π_1), constant values of P_e correspond to straight lines on a ROC diagram as shown in Figure 6. It then follows that the straight line defined by the minimum value of the probability of error, $P_e \geq P_e^{min}$, partitions the ROC diagram into two regions corresponding to achievable and unachievable performance. The triangular region in the ROC diagram inaccessible to ATR performance can be written as

$$R_{\pi_0\pi_1} = \{(P_{10}, P_{01}) \mid P_{01} \leq -P_{10}\frac{\pi_0}{\pi_1} + \frac{P_e^{min}}{\pi_1}\} \tag{10}$$

for given prior probabilities, (π_0, π_1), and is illustrated in Figure 6.

If the prior probabilities, (π_0, π_1), are varied in a parametric sense, both the slope of the line and the intercept in Equation 9 change producing a new region of inaccessibility. The set of all such regions, define a family of bounds on ATR performance.

A family of bounds can be generated from Fano's inequality by parametrically varying the values of (π_0, π_1) in Equation 10. For clarity, we examine these bounds in the context of a specific binary hypothesis testing problem. We start by defining the conditional probabilities

$$Pr(y \mid c_0) \equiv P_0(y) = \left(\frac{15}{16}, \frac{3}{64}, \frac{3}{256}, \frac{1}{256}\right), \tag{11}$$

and

$$Pr(y \mid c_1) \equiv P_1(y) = \left(\frac{1}{256}, \frac{3}{256}, \frac{3}{64}, \frac{15}{16}\right). \tag{12}$$

The Blahut bound is computed from Equation 5 by treating r as an independent variable and computing P_{01}. For this problem $c = 5.165$ and $\sigma^2 = 1.708$. This calculation yields only a portion of the bounding curve due to the restriction that $r > c$. To construct the other portion, the roles of the error probabilities P_{01} and P_{10} are reversed, and the computation repeated.

We illustrate such a family of bounds for the specific hypothesis testing problem introduced above. Figure 7 shows a family of bounds generated from multiple applications of Fano's inequality and Equation 10. For comparison, the Neyman-Pearson ROC curve and Blahut's bound are also presented.

Equation 10 introduced the concept of a region of inaccessiblity for the performance of an ATR system. If we consider the *union* of the regions of inaccessiblity over (π_0, π_1),

$$\mathcal{R} = \bigcup_{\pi_0\pi_1} R_{\pi_0\pi_1}, \tag{13}$$

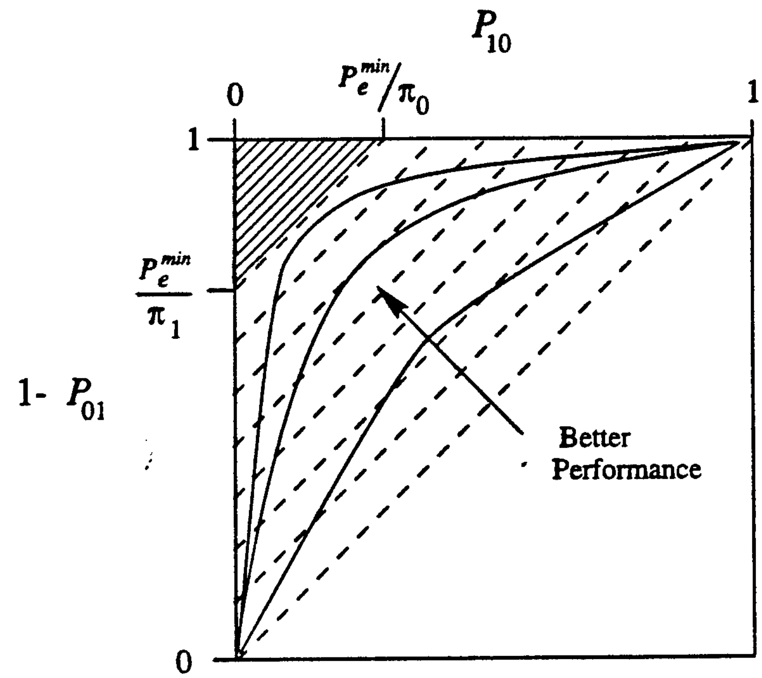

Figure 6: ROC diagram partitioned by a lower bound on the total probability of error P_e.

then it is evident that $\mathcal{R}$ itself defines a bound on ATR performance over all values of (π_0, π_1). This bound is readily seen to be described by the envelope formed by the sequence of lines generated by the family of Fano derived bounds. Hence, we call this new bound the Fano-Envelope bound.

3.5 Discussion

We have developed a family of bounds on ATR performance based on Fano's inequality from information theory and expressed in terms of the ROC diagram from hypothesis testing. We were able to demonstrate that the associated Fano-Envelope bound that is generated by the envelope of the family of bounds we obtained is a much tighter bound relative to the Neyman-Pearson ROC curve than the Blahut inequality.

¿From a pragmatic viewpoint, however, the requirements for computation of the bounds derived from Fano's inequality equal those for directly generating the exact ROC curve of a given ATR system.

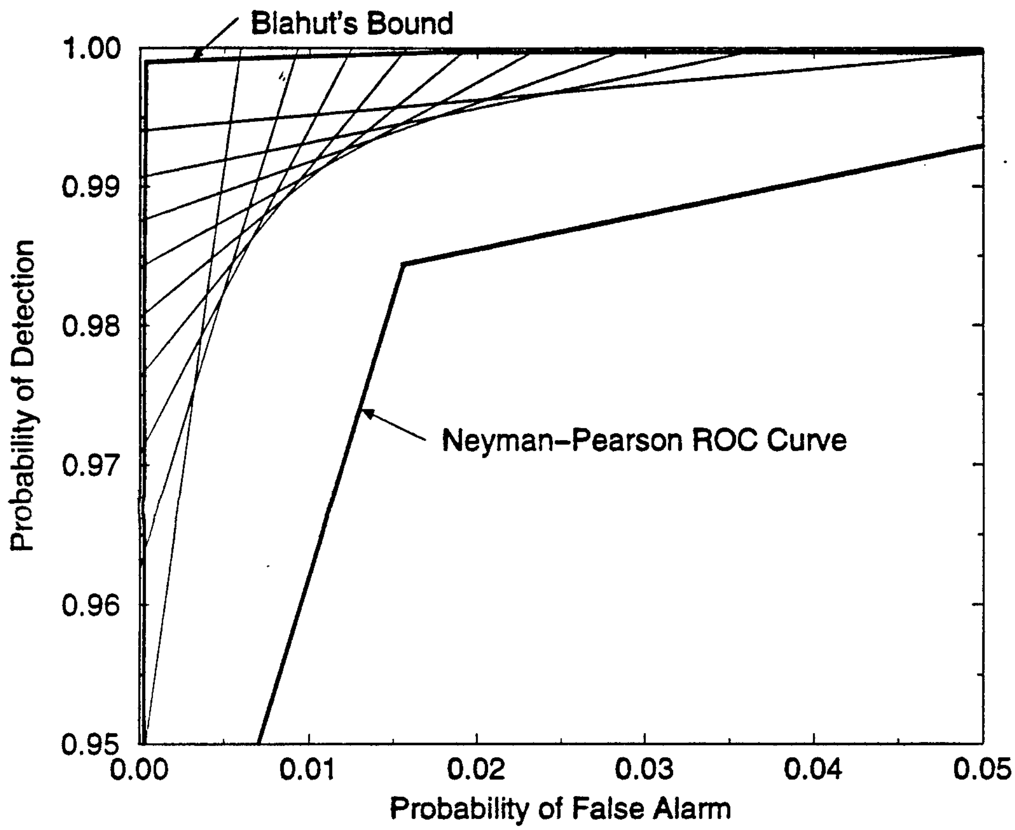

Figure 7: A family of bounds obtained from Fano's inequality. The Neyman-Pearson ROC curve and Blahut's bound are shown for comparison.

4 The Information Theoretic Decision Tree ATR System

4.1 Problem and Approach Description

4.1.1 Automatic Target Recognition (ATR) problem

The problem we are trying to solve in the following formulation of an ATR architecture can be stated as follows.

Given a set of images which represent some object at different poses (that is, a set of exemplars), construct a system capable of performing the following recognition task:

For a given image, perhaps containing a target related to the training suite of images by corruption with noise or obscuration and embedded on background clutter, estimate the pose of the most likely target and return a metric which describes (in some sense) the strength of confidence of the presence of the target at that pose in the image.

The brute-force approach to solving this problem would involve somehow comparing the given test image to each image in the training set of images (target suite database) and make a decision based on a similarity describing metric value.

However, ATR systems must make a decision in a short period of time. The database for a realistic training suite could contain hundreds of thousands (or even billions) of high-resolution images owing to the presence of several degrees of target pose, illumination, historical excitation and articulation freedom; thus brute-force methods can be unusably slow.

4.1.2 Decision Tree Approach

The basis for a *Decision Tree* approach involves creation of a set of *decision rules* for a given database and replacement of brute-force comparisons between target image and database images with testing of the target image against a contextually selected sequence of these decision rules. The goal is to make the operation of testing against these decision rules fast and to keep the total number of decision rules tested, to obtain a final decision, small.

Although the price for faster on-line processing will be time needed to construct a set of decision rules for a given set of database images, this can be done off-line, i.e., prior to the recognition process.

4.1.3 A Linear Decision Tree Architecture

Let's assume that there are K images in the training image database and each is $n_1 \times n_2$ pixels in extent, thus we can think of the suite of images as K vectors or points in the N-dimensional space that contains all possible images, where $N = n_1 \times n_2$.

As is well known, the problem of target recognition now corresponds to the division of this space into regions which describe equivalence classes. In the simplest, binary hypothesis case, we would assign membership of every point in the space to one of two equivalence classes,

that of H_0 which represents the no-target-present hypothesis and H_1 which represents the target-present hypothesis.

Generally, we would want to divide the space further into equivalence classes representing each possible target (multi-hypothesis problem) or representing values of a parameter associated with the target (composite hypothesis problem.) In the sequel we will treat the case of a single target type, with a continuous parameter represented as a set of discrete values (for which exemplars of the resulting target image are available in our training suite.) Furthermore, we shall concern ourselves with the division of the space into equivalence classes only with respect to the parameter values. That is, there will be no equivalence class associated with the H_0 hypothesis.

What would appear to be an odd rejection, that is, not considering the separation of the image space for purposes of the target detection, makes more sense when considering the operation of the classic likelihood ratio based Neyman-Pearson composite hypothesis test. In effect, we wish to maximize the likelihood ratio with respect to all possible parameter values. Then, a separate decision may be applied to the result regarding the target-presence hypothesis. Likewise, we shall construct our equivalence class division so as to derive the match metric between the given image and the most likely parametric representation. This final metric may then be judged on the basis of a contextual threshold value determined by a constraint such as probability of false alarm versus probability of detection. Incorporation of the final detection into the equivalence class construction would yield a complicated variation of class boundaries based on the constraints and associated thresholds rather than the fixed boundaries yielded by the current approach.

4.2 Hyper-plane Based Approximation

The construction of the equivalence class boundaries described above proceeds directly from the association of every point in the N-D image space with the exemplar that maximizes the a posteriori probability function. However, in general, these boundaries are surfaces with arbitrary curvature. That is, the expense of storing representations of these boundaries and of testing equivalence set membership of a point with respect to these boundaries is immense. In fact, exact representation in the most general case requires infinite data storage capacity since the boundaries are generally not describable by less than an infinite number of power series expansion coefficients. Hence, we need to consider from the onset the approximate representation of the decision boundaries.

One approximate representation has particular appeal and will the basis for the remaining development. Any continuous surface can be approximated to a any given degree by a sufficiently large set of planar facets. Furthermore, testing a point for inclusion inside a space defined by a set of L planes involves only computation of the inner products of the the point vector representation with respect to the L normals of the planes. Thus, the degree of approximation and computational complexity of the decision are both simply linked to the number of planes chosen for the representation.

Thus, building a set of decision rules can be associated with construction of hyper-planes which partition N-dimensional space into hyper-plane bounded domains, or polytopes, each containing only one exemplar database point. The minimal such construction, that is, one in which this segregation of each exemplar into a unique polytope is achieved with the

smallest number of decision planes will be the subject of the next phase of this development. As will be seen, we will have to moderate the minimalist nature of the construction to obtain a useful approximation of the original equivalence region. However the construction of significantly better approximations of the true equivalence class region based directly on geometric approximation of the true decision region is considerably more complicated and can be circumvented to some extent as will be discussed later.

4.2.1 Hyper-plane Decision Tree System

Construction of a minimal hyper-plane based division of our exemplars would appear to be related to a binary tree representation of the exemplar set. That is, considering each hyper-plane as a means to divide a set of points, the smallest number required to divide that set would be the number of nodes in a balanced binary tree which represents consecutive divisions of the exemplar set. Thus given K exemplars we would require on this basis no more than a binary tree with depth $log_2 K$ and $K-1$ associated decision planes. While the number of decision planes is large, the number of tests that must be conducted to ascertain the equivalence class membership is small, that being $log_2 K$.

As we discuss the construction of the hyper-planes that represent the decisions at the nodes of this binary tree, we will see that for our problem, $log_2 K$ decisions is a lower bound and not an entirely appropriate representation given the need to preserve performance within this approximate decision boundary representation.

The above statement can be motivated with the following observation. Consider an arrangement of image points in which division of the set into two equal halves (a prescription for the first node in a balanced binary tree) would require that the decision boundary pass immediately alongside a particular point. It should be obvious that attaining a minimal representation in this case will also lead to an incorrect association of that point with the nearest neighboring equivalence class nearly half the time given even a small noise induced distribution of image points about the exemplar point. Our minimization of representation must be tempered by issues related to performance in the context of stochastic perturbation of image points from the exemplar locations.

In the following we will derive an appropriate basis for the division of our exemplars by hyper-planes through an application of an information theoretic measure which does not result in a balanced binary tree.

The first-level hyper-plane, in a decision tree based scheme thus should, it would seem, divide all points into roughly, but not necessarily exactly, two equal subsets.The second-level hyper-planes divide two first-level subsets further roughly in halves again, and so on until all points are separated from each other or in other words until each point is *boxed* by a few hyper-planes (Fig.8).

Once such a system is constructed, the recognition process becomes nothing but a sequence of decision steps, where at each step the target image is tested with respect to which side of the hyper-plane, associated with the given node in the decision tree, it resides. At each step the number of possible equivalence classes that may be associated with the image under test is reduced approximately by half - until we reached a polytope containing only one exemplar point; this exemplar is then declared as the closest exemplar to the test image.

The system of hyper-planes has a binary tree structure and therefore the on-line complex-

Set of points and sample hyperplane division process

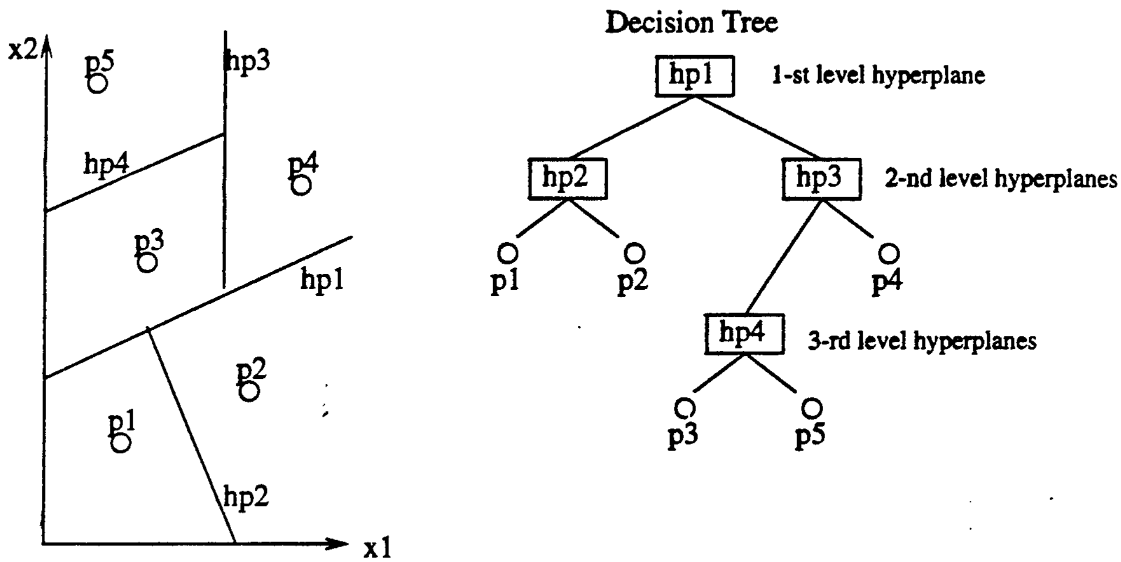

Figure 8: *Decision Tree Construction*

ity of forming the complete decision may be expected to be of the order $\log_2 K$ operations as opposed to K operations for the brute-force comparison approach.

4.2.2 The Information Theoretic Decision Tree

The simple assumption behind the approach outlined above is that in the high-dimensional image space, the boundaries of the decision region polytopes containing single points from the exemplar database are sufficiently far enough away from the exemplar points to allow the decision process to tolerate reasonable test image perturbation without causing a choice of a neighboring class. In other words we assume that given a target image derived from an exemplar image, say, $R = p_k + n$, then the noise, n, will not shift the test image point, R, into a polytope associated with an adjacent exemplar. Thus segregation of the exemplar points into individual polytopes cannot be the only basis for the placement of decision planes. Rather, we must consider the probability densities that are associated with the a priori statistics of the image generation process as these determine the appropriate placement of decision planes between adjacent exemplar points.

In the following sections we will develop a means for determining the placement of the decision planes, and implicitly, the form of the binary tree, so that a hierarchical decision process results. While other coarse-to-fine hierarchical decision processes have been proposed in the past, the approach to be described more fully below has the following special and attractive attributes:

- An information theoretic measure is used to drive the process of dividing the decision space. This optimizes the amount of "information" yielded by each decision made in the final process. If the imagery divides "cleanly" then this procedure guarantees the decision tree has the well known optimum number of decisions associated with that

number of image points. Otherwise, the decision process has the smallest number of decisions allowed by the data and consistent with minimal loss of discriminatory information at each step.

- The original image space is the setting for the decision process. Thus no artificial reductions in dimensionality are introduced and all relevant image information is preserved at each step in the process.
- The solution for the optimum decision tree is intractable, however, the penalty for using a sub-optimum solution is only additional decisions (hence processing time) and not any loss of ultimate estimation and target recognition performance.

4.3 Information Theoretic Formulation

In this section the mathematical background for the previously described concepts will be presented. Suppose now that for a given set of database points we have already placed a trial first-level hyper-plane. The following figure describes the current decision process and associated variables.

(Fig.9):

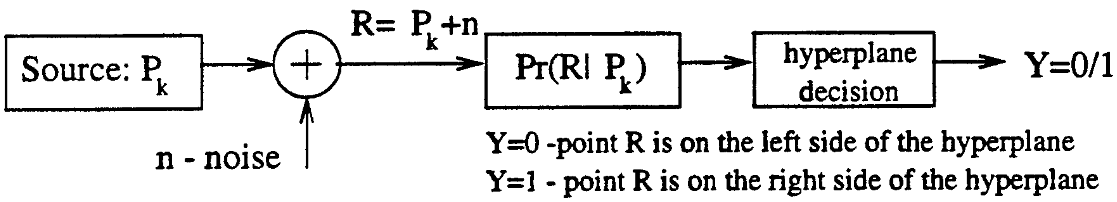

Figure 9: *First-level decision sub-system*

For a given test image R, which is assumed to be derived from one of the exemplar images $X = p_k$ corrupted with noise n, a decision, Y, is produced, which indicates on which side of the hyper-plane point R is located, hence an association of R with one of two groups of equivalence classes.

To determine the appropriate placement of the hyper-plane, consider a decision Y for some given hyper-plane placement: how much information I about the random hypothesis process X can we obtain based on observation of Y?

An appropriate definition of I in a decision theoretic context is the mutual information between Y and the random process X. Given this definition it is at once obvious that

$$0 \leq I = I(p_k; Y) \leq 1 \tag{14}$$

since we can derive no more than one bit of information from one binary decision. Obviously, we would want to derive as much information from such a decision as is possible. Thus, to make a well informed final decision from small numbers of intermediate decisions, we are motivated to solve for hyper-plane placement on the basis of maximizing the information derived from each intermediate decision.

It is important to note that what we are engineering is no improvement upon ATR performance over that available with a Bayesian or Neyman-Pearson construction. In fact,

there will in general be a loss of performance with respect to these provably optimal approaches. However, this development focuses on the minimization of computational effort (the decisions) while retaining good performance. This minimization is achieved by fixing a computational architecture and applying information theoretic optimization to minimize the effort required to derive information from the data.

4.3.1 Mutual Information Function

Now we must determine the best hyper-plane position so that the information measure, I, is maximized for a given set of exemplar points. In order to solve this problem we need to express I in terms of hyper-plane characteristics and the set $\{p_i\}$

Lets express I in terms of entropy [1] :

$$I(\{p_i\};Y) = \mathcal{H}\ (\{p_i\}) + \mathcal{H}\ (Y) - \mathcal{H}\ (\{p_i\},Y), \tag{15}$$

where

$$\mathcal{H}\ (x) = -\sum_{x_i} Pr(x = x_i) \log_2(Pr(x = x_i)) \tag{16}$$

Thus

$$\mathcal{H}\ (Y) = -\sum_{y_i} Pr(Y = y_i) \log_2(Pr((Y = y_i)) \tag{17}$$

$$\mathcal{H}\ (p_i) = -\sum_{p_i} Pr(p_i) \log_2(Pr(p_i)) \tag{18}$$

Note that $\mathcal{H}\ (p_i)$ is just a constant for a given probability distribution associated with the p_i.

$$\mathcal{H}\ (p_i, Y) = -\sum_{p_i}\sum_{y_i} Pr(p_i, Y = y_i) \log_2(Pr(p_i, Y = y_i)) \tag{19}$$

$$Pr(p_i, Y = y_i) = Pr(p_i) \int_{Y=y_i} p(r|p_i)dr \tag{20}$$

where
$Pr(p_i)$ is the probability of the occurrence of p_i,
$p(r|p_i)$ is the probability distribution function of the target r given pose p_i, and
$Pr(Y = y_i) = \sum_{p_i} Pr(Y = y_i|p_i)Pr(p_i)$.

Now, any hyper-plane in N-dimensional space can be uniquely defined by (N+1) coefficients. (For example the equation of a hyper-plane in 2 dimensions, a line, is given by $ay + bx + c = 0$, where $\{a, b, c\}$.) We will describe the hyper-plane by coefficients $w_0..w_N$.

Now, given hyper-plane coefficients $w_0..w_N$ we can uniquely compute Y for a given point r in the N-dimensional space, thus we can express I in terms of $w_0..w_N$ and the set of $\{p_i\}$.

The maximization of I with respect to $w_0..w_N$ is a highly non-linear global optimization problem for which possibly many suboptimal solutions can be found. In the current implementation of this work this optimization is performed by a multidimensional conjugate gradient method [11].

4.3.2 Decision Tree Construction

Given a procedure that places a hyper-plane to divide a given set of points so as to maximize the information gained by the decision, the decision tree construction process becomes nothing but recursive application of that procedure to the each subset of points derived by each such division. The hyper-plane coefficients are then stored and associated with nodes of the decision tree.

5 Directions for Further Research

The ITDT ATR architecture developed in the course of this work serves as a demonstration of the direct applicability of notions from information theory to realization and optimization of ATR technology. While the prototype has proven to be quite interesting and powerful, many directions for investigation and enhancement are open, including:

- implementation of 2-DOF, 3-DOF, ..., N-DOF versions;
- extensive testing with "real" data sets with detailed statistical characterization of sensor, target, environment, variation of several DOFs and high accuracy truthing;
- development of means to deal with the size of the decision tree storage, via specialized compression techniques;
- development of theory giving rise to a metric function that supports partial object recognition within this framework;
- implementation of a distributed network computer based system for the time consuming off-line computation of the decision tree solution.
- exploration of the performance and storage requirements associated with an ITDT which foregoes target orientation analysis (yielding only a target/no-target decision).

6 List of Publications

None to date.

7 Personnel Supported

1. Prof. David Cyganski
2. Prof. Richard Vaz
3. Ph.D. student James Kilian
4. Ph.D. student Sergey Perepelitsa

References

[1] T.M. Cover, J.A. Thomas "Elements of Information Theory," *Wiley Series in Telecommunications,* 1991, New York.

[2] R.E. Blahut, "Hypothesis Testing and Information Theory," *IEEE Transactions on Information Theory*, IT-20, 1974, pp. 405-417.

[3] H.V. Poor and S. Verdu, "A Lower Bound on the Probability of Error in Multihypothesis Testing," *IEEE Transactions on Information Theory*, IT-41, 1995, pp. 1992-1994.

[4] T.S. Han and S. Verdu, "Generalizing the Fano Inequality," *IEEE Transactions on Information Theory*, IT-40, 1994, pp. 1247-1251.

[5] R.M. Gray, "Entropy and Information Theory," *Springer-Verlag*, 1990, New York.

[6] R.G. Gallagher, "Information Theory and Reliable Communications," *John Wiley & Sons*, 1968, New York.

[7] D. Cyganski, R.F. Vaz, and C.R. Wright, "Model-Based 3-D Object Pose Estimation from linear image decomposition and direction of arrival techniques," *Proc. SPIE Proceedings, Conference on Model-Based Vision, vol. 1827*, November 1992, Boston. Viewable on the WWW, `http://xfactor.wpi.edu/~Works/Papers.html`.

[8] D. Cyganski, B. King, R.F. Vaz, and J.A. Orr, "ROC Analysis of ATR from SAR images using a Model-Based Recognizer Incorporating Pose Information" *SPIE 1995 Symposium on OE/Aerospace Sensing and Dual Use Photonics*, April 1995, Orlando, Florida. Viewable on the WWW, `http://xfactor.wpi.edu/~Works/Papers.html`.

[9] B.K. Hill, D. Cyganski, and R.F. Vaz, "SAR ATR Via Pose-tagged Partial Evidence Fusion," *AeroSense, SPIE International Symposium on Aerospace/Defense Sensing and Controls*, April 1996, Orlando, Florida. Viewable on the WWW, `http://xfactor.wpi.edu/~Works/Papers.html`.

[10] W. R. Jachimczyk and D. Cyganski, "Enhancements of Pose-tagged Partial Evidence Fusion SAR ATR," SPIE 1997, Aerosense 1997 Orlando, FL. Viewable on the WWW, `http://xfactor.wpi.edu/~Works/Papers.html`.

[11] W.H.Press *et.al*, "Numerical Recipes in C," *Cambridge University Press,* (1988).